Math 141

Lecture 24: Model Comparisons and The F-test

Albyn Jones[1]

[1]Library 304
jones@reed.edu
www.people.reed.edu/∼jones/courses/141

Nested Models

Two linear models are **Nested** if one (the *restricted model*) is obtained from the other (the *full model*) by setting some parameters to zero (i.e. removing terms from the model), or some other constraint on the parameters.

We can compare nested models fit to the same dataset with the F test.

Example

```
# Full Model
Mfull <- lm(Y ~ X + W + Z + T,
               data = MyDataSet)

# Restricted Model
Mres <- lm(Y ~ X + W, data = MyDataSet )
```

Fitting the restricted model is equivalent to forcing $\beta_Z = \beta_T = 0$ in the full model.

Comparing Nested Models

The crucial question is whether the residual sum of squares for the restricted model (RSS_R) is substantially larger than the residual sum of squares for the full model (RSS_F).

R. A. Fisher worked out the distribution of a ratio of the two under the null hypothesis that the restricted model is correct, which typically corresponds to the statement that some parameters are zero.

As usual, this story depends on the residuals having a at least an approximately normal distribution.

The F-Test

Assuming model validity, the F-ratio (F is for Fisher, by the way)

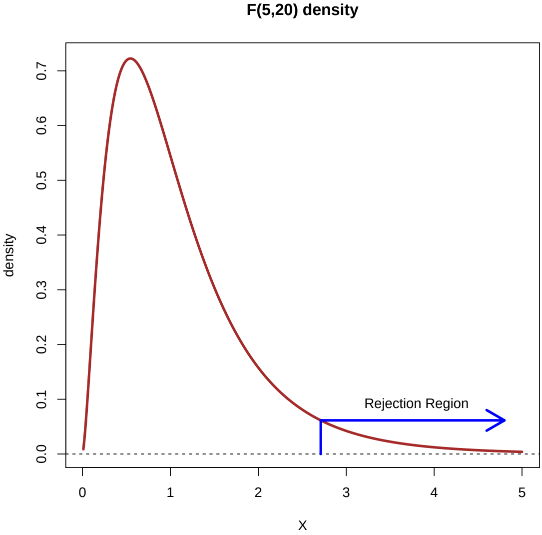
$$F_{df_N,df_F} = \frac{(RSS_R - RSS_F)/(df_R - df_F)}{RSS_F/df_F}$$

has an F distribution with degrees of freedom (df_N, df_F) if the restricted model is correct.
Note: $df_N = df_R - df_F$, and df_F and df_R are residual df from the two models.

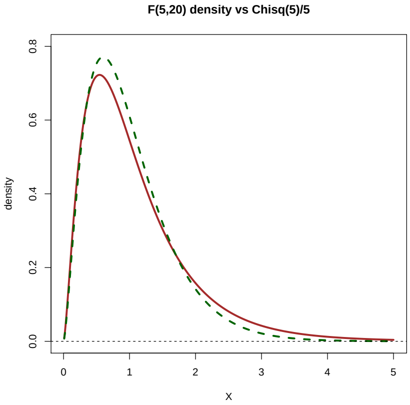
Reject: if $F > qf(.95, df_N, df_F)$

Note: $df_R - df_F$ is always the number of constraints on the parameters that converts the full model to the restricted model.

The F density



Analogy



$F_{k,n}$ is to χ^2_k as t_n is to $N(0,1)$. The denominator estimates σ^2. If we knew σ^2, the ratio would have a χ^2 distribution.

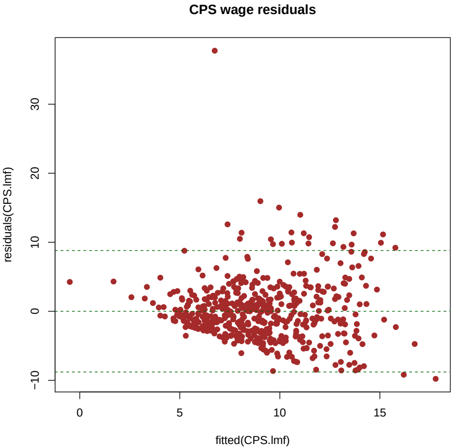
Connection to the t Distribution: $F_{1,k}$ is t_k^2

```
lm(formula = ht18 ~ ht2, data = Berkeley)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.1203    26.6572   1.205    0.233
ht2           1.5998     0.3031   5.278  2.2e-06
---

Residual standard error: 7.572
              on 56 degrees of freedom

F-statistic: 27.86 on 1 and 56 DF, p-value: 2.2e-06

> 5.278^2
[1] 27.85728
```

Example, CPS wage data summary

```
Call:
lm(formula = wage ~ race*sex + educ + age + union,
                             data = CPS)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.09772    1.46620  -4.841  < .001
raceW        0.06191    0.89024   0.070  0.94458
sexM         0.59693    1.10624   0.540  0.58970
educ         0.82717    0.07405  11.170  < .001
age          0.10481    0.01672   6.268  < .001
unionUnion   1.59479    0.51016   3.126  0.00187
raceW:sexM   1.77023    1.17363   1.508  0.13207
```

Plot Residuals!

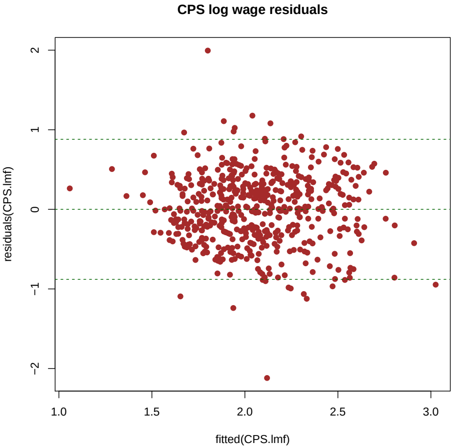


What Next?

Example, CPS log(wage) data summary

```
Call:
lm(formula = log(wage) ~ race*sex + educ +
                         age + union, data = CPS)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.330807   0.147882   2.237   0.0257
raceW       0.033692   0.089791   0.375   0.7076
sexM        0.112103   0.111577   1.005   0.3155
educ        0.085200   0.007469  11.407   < .01
age         0.011585   0.001686   6.869   < .01
unionUnion  0.221588   0.051455   4.306   < .01
raceW:sexM  0.133519   0.118374   1.128   0.2599

Residual standard error: 0.4446 on 527 df
```

Plot Residuals Again!



Better?

Looking for a parsimonius model?

None of the coefficients for race, sex, and the race*sex interaction were statisticially significantly different from zero. Let's fit a restricted model, dropping those non-significant explanatory variables.

Example, CPS log(wage) Restricted Model

```
Call:
lm(formula = log(wage) ~ educ + age + union,
                         data = CPS)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.511462   0.127785   4.003   < .01
educ        0.084898   0.007694  11.034   < .01
age         0.010742   0.001728   6.217   < .01
unionUnion  0.260260   0.052135   4.992   < .01

Residual standard error: 0.4593 on 530 df
```

Model Comparison

```
> anova(CPS.loglmr,CPS.loglmf)
Analysis of Variance Table

Model 1: log(wage) ~ educ + age + union
Model 2: log(wage) ~ race*sex + educ + age + union
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    530 111.81
2    527 104.16  3    7.6478 12.898 3.836e-08
```

Say WHAT? None of the omitted coefficients were statistically significantly different from 0! How can this happen?

The Null and Alternative Hypotheses

What is H_0?

The restricted model is correct. Informally: the restricted model fits as well as the full model.

Formally:

H_0: coefficients for the omitted terms are all 0.

Formally:

H_1: at least one omitted coefficient is not zero.

Important!

Individual t-tests are testing a null hypothesis for a single coefficient

$$H_0 : \quad \beta = 0$$

given we have controlled for the other variables in the model!

What was missing?

```
formula = log(wage) ~ sex + educ + age + union,
                        data = CPS)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.352998   0.126995   2.780  0.00564
sexM        0.228555   0.039400   5.801  < .001
educ        0.085473   0.007468  11.445  < .001
age         0.011727   0.001686   6.957  < .001
unionUnion  0.210191   0.051331   4.095  < .001
```

What happened?

The race*sex interaction was a distraction!

```
> cor(sex=="M",sex=="M" & race=="W")
[1] 0.8638632
```

Strongly correlated explanatory variables can be distractors, each does part of the work of predicting the response, neither seems important when the other is included.

Interpretation

The coefficient for the dummy variable for Males was about .23. What does that mean?

All other factors held equal, the difference between $log(wage)$ for males and $log(wage)$ for females is .23:

$$log(W) = OtherStuff + .23 \cdot sexM$$

Therefore

$$W = e^{OtherStuff + .23 \cdot sexM} = e^{OtherStuff} e^{.23 \cdot sexM}$$

The dummy variable sexM is 1 for males and 0 for females, so the difference is the multiplicative factor

$$e^{.23} \approx 1.26$$

Conclusion: Males with the same education level, age and Union status get paid about 26% more than corresponding females with the same covariate values.

R will try to prevent silliness

```
> anova(CPS.loglmr,CPS.lmf)
Analysis of Variance Table

Response: log(wage)
           Df  Sum Sq Mean Sq F value Pr(>F)
educ        1  21.481 21.4807 101.821 < .001
age         1   9.898  9.8976  46.916 < .001
union       1   5.257  5.2573  24.920 < .001
Residuals 530 111.811  0.2110
---

Warning message:
In anova.lmlist(object, ...) :
  models with response "wage" removed because
  response differs from model 1
```

Michelson's Data, full model

```
> summary(MF)
Call:
lm(formula = Speed ~ Run, data = Michelson)

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 299909.00      16.60 18067.739 0.000
Run2           -53.00      23.47    -2.258 0.026
Run3           -64.00      23.47    -2.726 0.007
Run4           -88.50      23.47    -3.770 0.000
Run5           -77.50      23.47    -3.301 0.001
```

Michelson's Data, restricted model

```
> Run1 <- Michelson$Run == 1

> summary(MR)

Call:
lm(formula = Speed ~ Run1, data = Michelson)

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) 2.998e+05  8.283e+00 36197.63  < 2e-16
Run1TRUE    7.075e+01  1.852e+01     3.82 0.000234
```

Model Comparison!

```
> anova(MR,MF)
Analysis of Variance Table

Model 1: Speed ~ Run1
Model 2: Speed ~ Run
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     98 537935
2     95 523510  3     14425 0.8726 0.4582
```

What was H_0, and what do we conclude?

Summary

The F test compares nested models fit to the same dataset.

It allows us to test hypotheses involving multiple parameters simultaneously.

If you wish to conclude that a collection of coefficients are all zero, or none of a subset of your explanatory variables predict the response, an F-test is the appropriate tool.